Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220448

Toward Automatically Identifying Legally Relevant Factors

Morgan GRAY [a,1], Jaromír ŠAVELKA [c] Wesley OLIVER [d] and Kevin ASHLEY [a,b]
[a] *Intelligent Systems Program, University of Pittsburgh, USA*
[b] *School of Law, University of Pittsburgh, USA*
[c] *School of Computer Science, Carnegie Mellon University, USA*
[d] *School of Law, Duquesne University, USA*

Abstract. In making legal decisions, courts apply relevant law to facts. While the law typically changes slowly over time, facts vary from case to case. Nevertheless, underlying patterns of fact may emerge. This research focuses on underlying fact patterns commonly present in cases where motorists are stopped for a traffic violation and subsequently detained while a police officer conducts a canine sniff of the vehicle for drugs. We present a set of underlying patterns of fact, that is, factors of suspicion, that police and courts apply in determining reasonable suspicion. We demonstrate how these fact patterns can be identified and annotated in legal cases and how these annotations can be employed to fine-tune a transformer model to identify the factors in previously unseen legal opinions.

Keywords. Automatic Text Identification, Multi-label Classification, Sentence Classification, Totality of the Circumstances Test

1. Introduction

We are investigating legal cases that assess the constitutionality of police decisions to search an automobile for drugs or to detain it for a canine search. Drug interdiction automobile stops have led to thousands of cases at the state and federal level, including the U.S. Supreme Court. Courts assess whether police had reasonable suspicion or probable cause to believe drugs are present in a car. With reasonable suspicion, an officer can briefly detain a motorist until a drug dog confirms or dispels the officer's suspicion. With probable cause, an officer may search the car. Officers can consider anything that enhances their suspicion of motorists, but officers frequently identify factors such as rental cars, strange travel plans, and the presence of strong air fresheners as increasing the likelihood that drugs will be found. These and others are referred to as "factors of suspicion." Officers must then decide whether they have probable cause to search the car or the lesser standard of reasonable suspicion to detain it until a drug dog arrives. If a search is conducted and something illegal is discovered in the search, only then will a judge determine whether there was adequate legal suspicion for detention.

[1]Corresponding Author: Morgan Gray, Learning Research and Development Center: 3420 Forbes Ave. Pittsburgh, PA 15260, USA; Email: mag454@pitt.edu

The open-ended nature of these legal standards, the enormous number of decisions, and the delay in judicial review cause uncertainty in deciding if there is reasonable suspicion in auto stop scenarios. Judges, lawyers, and police officers cannot read all 40,000 published decisions and determine how, courts, in the aggregate, view each suspicious factor an officer identified in a drug interdiction stop. Processing large amounts of text, however, is exactly what machines can do very well.

We hypothesize that a methodology for automatically assessing reasonable suspicion in drug interdiction stops is possible. We apply text analytic methods to auto stop cases in order to answer the following research question: Can a computer program apply ML/NLP to learn automatically to identify auto stop case outcomes and the factors of suspicion that courts consider in assessing the legality of the stops?

In this paper we report our initial progress in developing a type system of factors of suspicion, annotating auto stop cases in terms of factors, computing the level inter annotator agreement, and automatically identifying factors in case texts. Our ultimate goals are to compute the weight courts assign to each of the factors in isolation, as well as in combination with other factors and to assess the likelihood that a court will find facts sufficient for a search, thus providing a metric for assessing the legal merits of the police stops. This could lead to an empirically-based definition of reasonable suspicion and probable cause, provide insights about the efficacy of such searches, and inform new policies and procedures to improve the accuracy of traffic stops and lessen the influence of implicit bias and its effect on minority citizens.

2. Related Work

Factors are stereotypical patterns of facts that tend to strengthen or weaken a plaintiff's argument in favor of a legal claim. [1]. Since their introduction in the HYPO program [2], factors have become a staple knowledge representation technique in AI and Law. See [3]. Computational models of case-based legal argument such as [4], [5] and [6] now employ factors in modeling arguments with legal rules, cases, and underlying values. Researchers have also made some progress in automatically identifying applicable factors in the texts of legal decisions. [7] trained a machine learning approach with case summaries to identify trade secret misappropriation factors in other summaries. Wyner and Peters [8] employed an annotation pipeline to extract information related to factors in trade secret law from the full texts of cases. [9] trained a model to identify such factors in full texts of cases.

In a promising approach [10] Branting and colleagues trained a machine learning program called SCALE (semi-supervised case annotation for legal explanations) to label text excerpts in WIPO (World Intellectual Property Organization) domain name dispute cases by applicable legal issues and factors. The ultimate goal is for SCALE to employ the issue and factor labels in explaining its predictions of outcomes of new cases in terms of reasons that legal professionals would understand. Beyond connecting case sentences to reasons, issue and factor tags could connect the case to computational models of legal argument like those mentioned above. A text analytic program like SCALE could identify the applicable issues and factors, and a case-based model, equipped with a database of cases annotated by SCALE, could assist in explaining and testing its predictions.

We expect that our approach in this work will improve performance in learning to identify factors in a new and factually more diverse legal domain. We believe that

automatically identifying factors of suspicion in drug interdiction auto stop cases is more difficult than identifying factors in the WIPO domain name cases of [10]. For one thing, auto stop cases involve greater stylistic diversity than WIPO domain name arbitration cases. Auto stop decisions are written by judges, not arbitrators, in trial and appellate courts from jurisdictions across the country. Auto stop cases also likely involve more diverse factual scenarios than WIPO cases. Unlike Branting, et al. we apply a transformer model to identify the factors of suspicion in the auto stop case texts. The language model, RoBERTa, [11] has been pretrained on vocabulary from an extensive text corpus. The process of training the model on our corpus of auto stop cases then fine tunes the model.

3. Data

The raw data used consists of legal opinions collected from the Harvard Law School Case Law Access Project (HCAP. https://case.law/.) In order to collect this data two sets of search terms were used. The first was ("reasonable suspicion" and "canine"). The second added the phrase ("drug interdiction") to the first set of search terms. These searches returned roughly 2,500 cases, of which a legal expert selected 211 cases having confirmed that they dealt with the legal issue of interest. These cases were processed into text files containing relevant metadata and the raw text of the majority opinion, and the raw text of minority opinions if available. Of the cases retrieved, 70 cases were from federal courts, 141 were from state courts, accounting for 67[2] unique jurisdictions. Cases across different jurisdictions can be used in this task because the legal issue and surrounding case law is almost identical if not identical from jurisdiction to jurisdiction. There were 182 legal opinions drafted by an appeals court, and only 29 drafted by a trial court.

Based on those annotations from our corpus of 211 cases, we learned that 57% state decisions concluded that suspicion was present while 43% concluded that it was not. Of the federal courts, we learned that 77% of the decisions determined that suspicion was present and 23% determined that it was not. Thus, the overall percentage of suspicion found was 63%, with 37% finding that it was not present. These processed cases are the basis for the annotation task described in the next section.

3.1. Annotation

The targeted data include the sentences describing factors officers rely on in concluding reasonable suspicion exists. When a court assesses if reasonable suspicion is present, it assesses if all of the officer's observations, taken together, warrant reasonable suspicion.

3.1.1. Type System

The type system contains 19 factors of suspicion, shown in the Table 1 under five boldface headings **Occupant Appearance or Behavior**, **Occupant Status**, **Travel Plans**, **Vehicle**, and **Vehicle Status**.[3] Each factor is associated with a number indicating the *broad* category into which a factor falls and a letter.

[2]The two most frequent jurisdictions, each comprising 12 cases, were the State of Texas and the United States Court of Appeals for the Tenth Circuit. The top ten jurisdictions accounted for 90 of the 211 cases.

[3]The factors were determined by legal experts based on litigation experience and reading and analyzing hundreds of legal opinions.

Table 1. Factor Type System

1 Occupant Appearance or Behavior	**2 Occupant Status**
1A Furtive Movement	2E Motorist License
1B Physical Appearance of Nervousness	2F Driver Status
1C Nervous Behavior	2G Refused Consent
1D Suspicious or Inconsistent Answers	2H Legal Indications of Drug Use
	2I Motorist's Appearance Related to Drug Use
3 Travel Plans	**4 Vehicle**
3J Possible Drug Route	4L Expensive Vehicle
3K Unusual Travel Plans	4M Vehicle License Plate or Registration
	4N Unusual Vehicle Ownership
5 Vehicle Status	**6 Other Annotation Labels**
5O Indicia of Hard Travel	6T Other
5P Masking Agent	6U Possibly Off Point
5Q Vehicle Contents Suggest Drugs	6V Suspicion Found? - No
5R Suspicious Communication Device	6W Suspicion Found? - Yes
5S Suspicious Storage	

Each category encompasses factors related to a particular topic. For example, Category 2 includes factors related to the status of the vehicle's occupant. Factor 2H, *Legal Indications of Drug Use* is an appropriate label for sentences describing a situation where a motorist has a prior conviction for a drug offense, or has an active warrant for a suspected drug offense. Thus, although the **Occupant Status** factors describe different facts, they all concern legal indications that an individual may be a drug user.

Category 6, **Other Annotation Labels** contains options used to annotate other important aspects of a legal opinion. The *Other* category should only be used if an annotator believes that sentence, which clearly describes a factor the court found to be suspicious, does not *reasonably fit into any defined type, i.e., types 1A-5S*. Annotators can use the *Possibly Off Point* category to indicate that the case was not relevant to the legal issue of whether reasonable suspicion existed to extend a traffic stop. The *Suspicion Found? - No/Yes* categories are to be used to annotate the legal conclusion reached by the court as to whether reasonable suspicion existed.

3.1.2. Annotation Task

We hired 6 law students and 1 recent law school graduate to annotate each factor in the corpus of 211 cases. All of the annotators had undergone more than one year of formal legal education at an accredited law school.

The annotators received 10–15 hours of training. They were introduced to the legal problem in some depth. We introduced them to the type system outlined in Table 1 with a lecture and a detailed discussion. A *Factor Glossary* provided specific descriptions and examples of each factor type. We introduced the annotators to *Gloss*, a convenient online annotation environment developed by Jaromir Savelka that supports annotating sentences by color-coded highlighting. [12] In addition to the *Factor Glossary*, the annotators employed an *Annotation Guideline* containing detailed instructions on how to annotate cases. The Guideline provides specific instructions as to the spans of text to (or not to) be annotated and the labels to be applied. Working in a group, the annotators worked

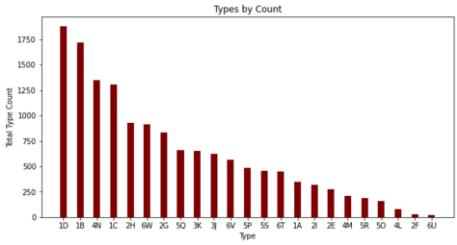

Figure 1. All Annotations: Counts by Type

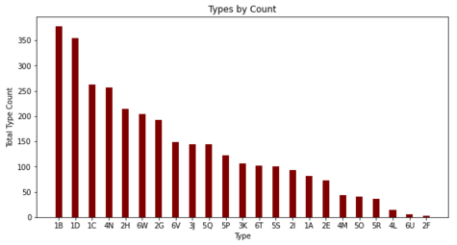

Figure 2. Gold Annotations: Counts by Type

through an initial case with the aid of a legal expert. The annotators then individually annotated five cases. The legal expert reviewed the annotations, corrected mistakes, and clarified points of misunderstanding. This continued for one more session with expert feedback before the students began annotating on their own.

All cases in the data set were annotated by at least two annotators. For the first 100 cases, both annotators were required to "resolve" any conflicts in their annotations. If they could not agree on an annotation, both annotations were kept. Finally, a legal professional reviewed all annotations and made his own judgment of how any disputed sentences should be labeled. The result served as the gold standard for the training data.

3.1.3. Annotation Outcome

The total number of annotations (i.e., annotations from all cases and all annotators) in the data set is 14,434.[4] The bar chart in Figure 1 shows the total number of annotations in the data set by type. The training data in the experiments consisted of 3,121 annotations. The breakdown of these annotations by type is show in Figure 2. The histograms appear to have a similar shape, and the bar for each type remained relatively in a similar position. This indicates that the overall frequency of annotations and the frequency in the gold standard annotations used for training data are similar.

The co-occurence matrix, which accounts for all annotations, reveals some interesting features of the data. The strong co-occurence between 6W, suspicion found, and 2G, the motorist denied consent to search is noteworthy. Under U.S. law, it is illegal to consider a motorist's refusal to consent as a factor making it more likely that drugs are present. Although officers may not be expressly rely on a motorist's refusal, refusing consent shows up frequently when suspicion is found and infrequently when suspicion is not found.

We employed Cohen's κ [13] metric to measure the agreement between two annotators. We measured inter-annotator agreement on all cases in the data set. The mean coefficient for the first 100 annotated cases was 0.544. The mean coefficient for the second 100 annotated cases was 0.601. The overall mean was 0.57. These scores indicate moderate agreement according to [14]. From these scores it appears that annotators improved with repetition and correction. Importantly, sentences labeled by the same factor typically do *not* describe the same facts. Different facts may fall into the same category, and similar facts may be described differently. Given the complexity of factor identification, moderate agreement seems reasonable.

[4] After annotations were assessed, cases that were to be determined to be off point were removed from the data set. Only five cases were removed on these or similar grounds.

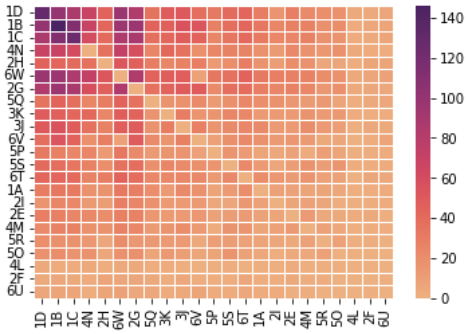

Figure 3. Co-occurrence Matrix by Type

4. Experiments

To assess whether automatic identification of factors in full text is feasible with this annotation scheme, we employed a pre-trained and subsequently fine-tuned language model. We classified the annotated types in order to distinguish among sentences that describe a factor, a conclusion, and non-typed sentences. Factor-related sentences were characterized by the applicable factors. Approximately 125,000 sentences in the data set were non-types and 14,000 cases in the data set represented a type. This is a multi-label classification problem in that sentences can describe more than one factor. For example:

> At the hearing, the officer testified that the reasons for the search were: Berry's nervousness, his uncertainty about whether his son was working or not, the fact that he was driving a rental car, the rental contract, Berry's looking down the interstate before answering some questions, Berry's failure to remember that his son lived in Decatur, the plastic garbage bag in the backseat, and the long trip from South Carolina only to stay a few hours.[5]

This sentence would be labeled in terms of the physical appearance of nervousness (1B), suspicious answers (1D), unusual vehicle ownership (4N), nervous behaviour (1C), suspicious storage (5S), and unusual travel plans (3K). Roughly 1,500 annotated sentences were annotated with one or more type. In order to deal with multi-labelled sentences, one-hot encoding was employed. The length of each encoding was 24, due to the 24 different possible types. An encoding representing one type was represented as follows:

$$[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$$

An encoding of the sentence shown above would look like this:

$$[0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]$$

The sentences were split into testing and training sets using a 60/40 training to testing split. The sentences and labels were then used to fine tune the roBERTa model over the course of 15 epochs, with evaluation occurring during training.

[5] Berry v. State, 547 S.E.2d 664 (Ga. Ct. App. 2001)

Table 2. Classification Report: Multilabel Classification

	precision	recall	f1-score	support
no_type	0.99	0.99	0.99	8381
1B Physical Appearance of Nervousness	0.92	0.89	0.90	62
2H Legal Indications of Drug Use	0.89	0.78	0.83	54
4N Unusual Vehicle Ownership	0.82	0.78	0.80	51
2G Refused Consent	0.79	0.73	0.76	30
3J Possible Drug Route	0.90	0.75	0.82	24
6W Suspicion Found? - Yes	0.92	0.83	0.88	42
5P Masking Agent	0.80	0.84	0.82	19
1C Nervous Behavior	0.78	0.85	0.81	53
3K Unusual Travel Plans	0.82	0.64	0.72	14
1D Suspicious or Inconsistent Answers	0.93	0.72	0.81	75
5Q Vehicle Contents Suggest Drugs	0.79	0.79	0.79	34
4M Vehicle License Plate or Registration	0.67	0.25	0.36	8
1A Furtive Movement	0.79	0.58	0.67	19
6V Suspicion Found? - No	0.71	0.87	0.78	31
6U Possibly Off Point	0.00	0.00	0.00	1
6T Other	0.79	0.58	0.67	19
2E Motorist License or Identification	1.00	0.91	0.95	11
2F Driver Status	0.00	0.00	0.00	0
5R Suspicious Communication Device	0.00	0.00	0.00	3
5S Suspicious Storage	0.73	0.84	0.78	19
5O Indicia of Hard Travel	0.00	0.00	0.00	6
2I Motorist's Appearance Related to Drug Use or Sale	0.92	0.85	0.88	13
4L Expensive Vehicle	0.00	0.00	0.00	3
micro avg	0.98	0.98	0.98	8972
macro avg	0.66	0.60	0.63	8972
weighted avg	0.98	0.98	0.98	8972
samples avg	0.92	0.92	0.92	8972

5. Results

The results of these experiments are shown in Table 2. Where the training data contained more than 10 instances of a type, the f-1 scores for classifying typed sentences ranged from 0.67 for factor 1A, Furtive Movement to 0.90 for 1B, Physical Appearance of Nervousness and 0.95 for 2E, Motorist License or Identification. Types that occurred in the test set fewer than eight times were not predicted or posted very low f-1 scores.

5.1. Discussion and Error Analysis

The results provide promising evidence that automatically identifying and extracting the relevant factors in legal opinions is feasible. On a number of occasions the classifier correctly identified a sentence that had been mislabelled as 'no type':

> "I don't have to let you search."

> "In determining the legality of a stop, courts do not attempt to divine the arresting officer's actual subjective motivation for making the stop; rather, they consider from an objective standpoint whether, given all of the circumstances, the officer had a reasonable and articulable suspicion of wrongdoing."

The first sentence in the above quote is describing the factual situation where an individual is refusing consent to search. Although the human annotators did not label it, the classifier correctly assigned the label of Refused Consent (2G). The second sentence describes a court's conclusion that there was not reasonable suspicion. Again, the human annotators missed it, but the classifier correctly assigned the appropriate label.

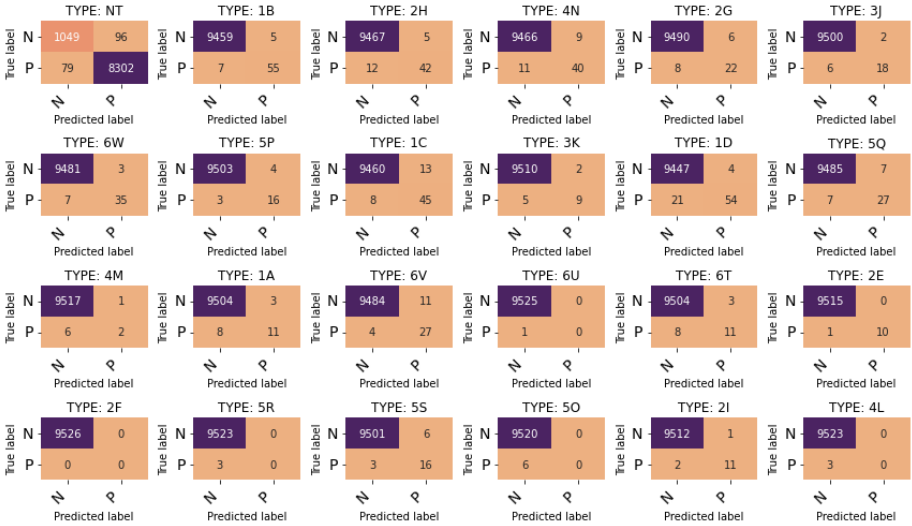

Figure 4. Confusion Matrix by Type

Whether precision and recall will be high enough for the intended downstream tasks is another question. As noted, we aim to compute weights for the factors across a large number of cases; high f-1 scores will be important.

In order to understand how we might improve performances, we undertook an error analysis. The confusion matrices are shown in Figure 4. For each matrix, the upper left quadrant reflects true negatives, and the bottom right represents true positives. For those factors that seemed to have higher numbers of errors, we examined some of the mislabeled instances to see if we could explain the errors.

The annotation guidelines require annotators to mark up only those sentences that indicate that a factor is present in the case. They were instructed *not* to annotate similar language in the opinion where the court discusses the legal significance of a factor generally or describes factors in other cases as courts may do in drawing comparisons across cases. Take, for example, these two sentences:

> Finally, Deputy Trammel testified that Mr. Powell appeared "excessively nervous" and remained so throughout the entire encounter, even when the deputy returned to Mr. Powell's vehicle to let him know the deputy was only giving him a warning citation. . . .
>
> Of course, even law-abiding citizens exhibit signs of nervousness when confronted by a law enforcement officer, and we have repeatedly held that "nervousness is of limited significance in determining reasonable suspicion."

Semantically, these sentences are similar. The first sentence should be annotated because it describes a fact present in the case at hand. The second sentence should not be annotated because the court is describing the legal treatment of nervousness as a factor. Given the semantic similarity of the sentences, the classifier would likely treat them the same. During one of the runs of the experiment described above, the classifier erroneously treated the first sentence as not belonging to a type.

Ignoring sentences that appear near legal citations may be a feasible path to eliminate at least some sentences that may confuse the classifier. As mentioned, some sentences describing a factor in another case may appear in a case at issue, for instance where a judge draws an analogy to or distinction from another case. Due to the nature of legal case citation rules, these semantically similar sentences will appear near case citations. It should be possible to filter sentences appearing close to a legal citation.

Another problem is the enormous variety of ways in which the facts characterized by a factor are expressed in the cases. The confusion matrix for Type 1D, Suspicious or Inconsistent Answers, shows a many false negatives. Upon inspecting some of these classification errors, we observed the following sentence:

> "While Anguiano attempts to highlight some of the consistencies in the men's stories, any general consistency cannot serve to dispel the contradictions of basic details, such as the name of the car's owner and the person for whom they were looking."

The sentence was properly labeled as a suspicious answer, however, the classifier predicted it as a not belonging to a type. Given the complexity of this example, we suspect that the high number of sentences properly labeled as 1D but misclassified by the model as 'no type' may be an example of this problem.

Another sentence describes a legal indication of drug use, factor 2H: having been convicted of a drug crime in the past.

> "Coupled with his observations of the items which, based on his training and experience, indicated narcotics trafficking, Officer MacMurdo was also very familiar with Defendant's prior criminal history and knew that Defendant was previously found in possession of drugs and a firearm."

The model erroneously predicted 'no type'. The example illustrates how differently the concept 'having been convicted of a drug crime in the past' can be expressed. We believe that increasing the quantity of annotated training data will assist the model to learn to correctly classify instances of factors like these.

Type 3J, Possible Drug Route, also had a relatively high number of false negatives. In the following sentence, we observed a different classification error also related to the variety of ways to express similar facts:

> "As he testified, it is commonly used to transport contraband throughout this state."

The sentence described a route known for drug trafficking, but the classifier predicted 'no type' when it should have predicted an instance of factor 3J. The anaphoric reference to 'it', a 'route' mentioned only in a previous sentence, probably interferes with a correct classification. Ambiguous pronoun references are likely to remain problematic.

We do aim to fix another common error in the data caused by problems with our sentence splitting algorithm [15]. It sometimes failed to correctly identify an annotated sentence. Essentially, the algorithm "missed" annotation, treating it as unannotated text. In future iterations we will improve the splitting algorithm to catch more annotations.

6. Conclusion

The results provide evidence that a program can learn to automatically identify factors in a new and factually more diverse legal domain, drug interdiction auto stop cases. In

future iterations of this work we intend to move from an off-the-shelf classifier to a model that has been specifically tuned for this task, for example by employing Legal-BERT, which has been pretrained on legal vocabulary from a large case law corpus. See [16]. Since the frequency of types appears to dramatically affect performance, we will collect more annotated data. For example, we are exploring how to integrate the annotation activities into pedagogical activities so that students can learn skills of close reading and legal argumentation as they annotate cases. Where we cannot increase the annotated data sufficiently, we will explore generating synthetic training examples for low frequency categories using resampling or similar techniques.

References

[1] Ashley KD. Artificial Intelligence and Legal Analytics. Cambridge U. Press; 2017.
[2] Ashley KD. Modeling Legal Arguments: Reasoning with Cases and Hypotheticals. MIT Press; 1990.
[3] Bench-Capon T. HYPO's legacy: introduction to the virtual special issue. Artificial Intelligence and Law. 2017;25(1):205-50.
[4] Grabmair M. Modeling Purposive Legal Argumentation & Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism. U. Pitt.; 2016.
[5] Chorley A, Bench-Capon T. An empirical investigation of reasoning with legal cases through theory construction and application. AI and Law. 2005;13(3):323-71.
[6] Chorley A, Bench-Capon T. AGATHA: Using heuristic search to automate the construction of case law theories. Artificial Intelligence and Law. 2005;13(1):9-51.
[7] Ashley KD, Brüninghaus S. Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law. 2009;17(2):125-65.
[8] Wyner A, Peters W. Towards annotating and extracting textual legal case factors. In: SPLeT-2012; 2010. p. 36-45.
[9] Falakmasir M, Ashley K. Utilizing Vector Space Models for Identifying Legal Factors from Text. In: JURIX 2017. vol. 302. IOS Press; 2017. p. 183-92.
[10] Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, et al. Scalable and explainable legal prediction. Artificial Intelligence and Law. 2021;29(2):213-38.
[11] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.
[12] Savelka J, Ashley KD. Segmenting US Court Decisions into Functional and Issue Specific Parts. In: JURIX; 2018. p. 111-20.
[13] Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.
[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977;33(1):159-74.
[15] Savelka J, Walker VR, Grabmair M, Ashley KD. Sentence boundary detection in adjudicatory decisions in the united states. Traitement automatique des langues. 2017;58:21.
[16] Zheng L, Guha N, Anderson B, Henderson P, Ho D. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: Proc. 18th Int'l Conf. on AI and Law; 2021. p. 159-68.